Range Results in XML Retrieval

Charles L. A. Clarke
School of Computer Science, University of Waterloo, Canada
claclark@plg.uwaterloo.ca

1. INTRODUCTION

To date, retrieval results for the INEX adhoc task have been restricted to simple XPath location paths with positional predicates, such as

```
/article[1]/bdy[1]/sec[5]/ss1[1]/p[4]
```

which specifies the fourth paragraph in the first subsection of the fifth section of the first body of the first article. This element of the document "`ex/2001/x3047.xml`" is an example of a high exhaustivity and specificity (3,3) element for INEX adhoc topic 165. Since elements are the standard unit of retrieval at INEX, the retrieval system must choose between this paragraph and the subsection that contains it. Unfortunately, the subsection may be too broad, but the paragraph may be too narrow.

XPath is considerably more expressive, and it may be fruitful to enlarge the class of possible results to include more of its features. Furthermore, I believe users would be better served by increasing the the set of potential results beyond single document elements. In particular, I believe it would be beneficial to express retrieval results as ranges of elements or text, for example as the "first three paragraphs in section 8," an approach which better reflects the way in which people informally describe portions of books and other documents. This short opinion paper marshalls a modest amount of evidence in support of range results, and briefly examines the availability of appropriate facilities in XPath to support these ranges.

2. RANGE RESULTS AT INEX

The potential benefits of range queries may be seen in the INEX 2004 adhoc relevance judgments. Of the 5229 elements judged as highly exhaustive and specific, at least 1700 (32%) are part of larger range of elements with identical tag names.

For example, the paragraph given above is part of the larger range of (3,3) elements

```
/article[1]/bdy[1]/sec[5]/ss1[1]/p[3]
/article[1]/bdy[1]/sec[5]/ss1[1]/p[4]
/article[1]/bdy[1]/sec[5]/ss1[1]/p[5]
/article[1]/bdy[1]/sec[5]/ss1[1]/p[6]
```


INEX Workshop on Element Retrieval Methodology, July 2005

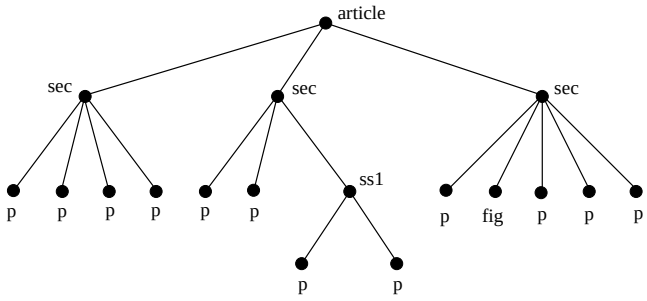

Figure 1: ***Example XML tree.***

By taking some liberties with XPath, this range of elements might be better expressed as

```
/article[1]/bdy[1]/sec[5]/ss1[1]/p[3 to 6]
```

It may be that this range is a more appropriate result than the individual paragraphs or the entire subsection — which is also a (3,3) element.

The inclusion of ranges in INEX retrieval results would necessitate changes to evaluation metrics and methodology, and the technique of assessing individual elements may have to be abandoned. Nonetheless, I believe it is feasible to extend the current INEX approach in a reasonable fashion, without introducing additional complexity. I outline one proposal next.

While struggling with the relevance assessment tools for INEX, I have often wished for a yellow highlighter that would allow me to directly select a section of a document for judging. Imagine a document marked up with a highlighter to indicate relevant regions. Potentially, each highlighted region could be labeled with exhaustivity and specificity attributes, and the relevance of an element (or a range of elements) could be determined from the attributes and proportion of highlighted text it contains. Moreover, if highlighting is permitted at the level of individual sentences and words, explicit labeling of specificity becomes unnecessary, since we may assume that only highly specific regions will be highlighted.

3. RANGE RESULTS IN XPATH

While we might informally speak of "the first three paragraphs" of a section, this statement can have many interpretations in the formal context of XML trees and XPath expressions. Consider the document in figure 1. While the

meaning of the statement is clear with respect to the first section, the same is not true of the second and third sections, since the third paragraph of second section is contained within a subsection and the first three paragraphs of the third section include a figure.

Straightforward attempts to specify "the first three paragraphs" in XPath further illustrate the ambiguities. For example, the expression

```
/article[1]/section[2]/p[position() <= 3]
```

will not include the first paragraph of the subsection. If it is correct to include this paragraph, an expression such as

```
/article[1]/section[2]//p[position() <= 3]
```

would be necessary. Similarly, the expression

```
/article[1]/section[3]/p[position() <= 3]
```

includes the first three paragraphs of the third section, but excludes the figure. Depending on the topic and the document, the inclusion of the figure may or may not be desired.

Other aspects of XML and XPath further complicate the specification of range results. Elements with the same logical type may receive different tag names. For example, in the current INEX adhoc test collection the tags "`p`" and "`ip1`" both indicate paragraphs. If ranges are to be accepted as retrieval results, then our method of expressing these ranges must accommodate this difference. On the other hand, a general ability to accept any XPath expression as a retrieval result is probably unneeded. A retrieval result that represents the second paragraph in every section

```
//sec/p[2]
```

is likely to have little value in a document-oriented context.

In XPath 2.0, it is relatively simple to express a range in a document as a pair of endpoints. Given two location paths, X and Y, the expression

X`/following::*[. <<` Y`]`

includes all the elements between them. In order to accept ranges as INEX retrieval results it may be sufficient to represent them as an (X, Y) pair. Using this representation, it would not be possible to exclude undesirable elements, such as the figure in third section of our example, but most of the potential benefits of range results could be realized.

4. SUMMARY

Constraining INEX results to single elements unnecessarily eliminates some of the potential benefits of XML retrieval, possibly forcing retrieval systems to return inappropriately narrow or broad results. Ranges are a natural means of specifying portions of documents and should be supported at INEX.

5. ACKNOWLEDGMENT

Thanks to Frank Tompa for his comments on this topic and his assistance with XPath 2.0.